

Post-processing through linear regression

B. Van Schaeybroeck and S. Vannitsem

Koninklijk Meteorologisch Instituut (KMI), Ringlaan 3, 1180 Brussels, Belgium

Received: 23 July 2010 – Revised: 15 December 2010 – Accepted: 8 February 2011 – Published: 7 March 2011

Abstract. Various post-processing techniques are compared for both deterministic and ensemble forecasts, all based on linear regression between forecast data and observations. In order to evaluate the quality of the regression methods, three criteria are proposed, related to the effective correction of forecast error, the optimal variability of the corrected forecast and multicollinearity. The regression schemes under consideration include the ordinary least-square (OLS) method, a new time-dependent Tikhonov regularization (TDTR) method, the total least-square method, a new geometric-mean regression (GM), a recently introduced error-in-variables (EVMOS) method and, finally, a “best member” OLS method. The advantages and drawbacks of each method are clarified.

These techniques are applied in the context of the 63 Lorenz system, whose model version is affected by both initial condition and model errors. For short forecast lead times, the number and choice of predictors plays an important role. Contrarily to the other techniques, GM degrades when the number of predictors increases. At intermediate lead times, linear regression is unable to provide corrections to the forecast and can sometimes degrade the performance (GM and the best member OLS with noise). At long lead times the regression schemes (EVMOS, TDTR) which yield the correct variability and the largest correlation between ensemble error and spread, should be preferred.

1 Introduction

Meteorological *ensemble* prediction systems provide not only a forecast, but also an estimate of its uncertainty. The ensembles are known to display some deficiencies

(Leutbecher and Palmer, 2008) which can be partially corrected through post-processing. Such post-processing consists of two steps. Firstly, regression is built between forecast and measurement, available during a certain training period, and secondly, the regression is applied to new forecasts. Regression methods are also of primary importance in the rapidly evolving research field concerning the combination of short-term multi-model climate forecasts (Van den Dool, 2006).

The classical linear regression approach of ensemble regression is based on ordinary least-square (OLS) fitting. This approach has some weaknesses which can be detrimental in the context of ensemble forecasts. In the present work, it is shown that other linear regression schemes exist which overcome them. One of the well-known problems with the classical linear regression approach is the fact that the corrected forecast converges to the climatological mean for long lead times (Wilks, 2006). Classical linear regression, therefore, fails to reproduce the natural variability caused by a progressive decrease of correlations between the true trajectory and the forecast data. To overcome this decrease of forecast variance, a few alternatives were already introduced. First of all, Unger et al. (2009), in an effort to prolong the correlation time, take a “best member” approach and average over the ensemble of forecasts to obtain the OLS regression parameters. In addition, they compensate the lack of climatological variability by a kernel method which consists in adding Gaussian noise, an approach also used by Glahn et al. (2009). In Vannitsem (2009), a new regression scheme was proposed which accounts for the presence of both the observational errors and the forecast errors. This approach which gives the correct variability at all lead times was tested against the non-homogeneous Gaussian regression using re-forecast data of ECMWF (Vannitsem and Hagedorn, 2011) and found to have good skill.

Here, we present a comparison of several linear regression methods to post-process ensemble forecasts. In addition, we



Correspondence to:
B. Van Schaeybroeck
(bertvs@meteo.be)

introduce two new regression schemes: a time-dependent generalized Tikhonov regularization method (TDTR) and a geometric-mean regression, analogous to the one presented in Draper and Yang (1997). Other schemes under consideration include the ordinary least-square method (OLS), the total least-square (TLS, Van Huffel and Vandewalle, 1991), the “best member”-OLS method (Unger et al., 2009) and the EV-MOS method, recently proposed by Vannitsem (2009). The latter is generalized for an arbitrary number of predictors and to cope with multicollinearity. The comparison is performed based on three criteria: a correct variability, a reduced forecast error and the ability to deal with multicollinearity.

The different regression schemes are tested in the context of the Lorenz 1963 model, focussing on the validity of the proposed criteria. We introduce both model and initial-condition errors and consider the dynamics of the statistical features of the corrected-forecast errors.

Section 2 details the problems associated with OLS for ensemble forecasting. The different regression approaches are then introduced in Sect. 3a–f and their quality is evaluated in the context of the low-order Lorenz model (1963) in Sect. 4. The ensemble skills are discussed in Sect. 5. Finally the conclusions are drawn in Sect. 6.

2 Linear regression and criteria

Consider a system for which a series of measurement data are available for a variable X , as well as one or more forecast models. When running model(s) multiple times using slightly perturbed initial conditions and starting at different dates, a set of forecasts is produced. The problem is how to optimally combine both past forecast and measurement data such as to extract as much information as possible and to correct future forecasts. We outline here different approaches to achieve this goal using linear regression.

Assume that the forecast data for P variables V_p ($p = 1, \dots, P$) is assembled in the $N \times P$ matrix \mathbf{V} ; here N is the number of ensemble forecasts multiplied by the number of ensemble members. Also consider the N measurement data for the variable X that are contained in the vector \mathbf{X} . Regression consists now of finding a solution for the P regression coefficients contained in the vector $\boldsymbol{\beta}$ such that:

$$\mathbf{X} \approx \mathbf{X}_C \quad \text{with} \quad \mathbf{X}_C = \mathbf{V}\boldsymbol{\beta}. \quad (1)$$

Here we call X_C the predictand or the corrected forecast while the variables V_p are the predictors. We take for the first predictor, V_1 , the corresponding model observable associated with X and V_1 is, therefore, also referred to as the uncorrected forecast. The near equality in Eq. (1) is achieved by minimizing some cost function, yet to be defined (see Sect. 3). Note that without loss of generality, we assume the mean value of all variables to be shifted to zero. Let us

define the error:

$$\varepsilon_X = X - \sum_{p=1}^P \xi_p \beta_p, \quad (2)$$

where ξ_p is the corrected predictor associated with V_p . The value ξ_p may be looked upon as the value of V_p after being deprived of errors of any sort. We denote the discrepancy between ξ_p and V_p as follows:

$$\varepsilon_{V,p} = V_p - \xi_p. \quad (3)$$

Note that the values of ξ_p are usually hidden and mostly introduced for optimization purposes. In order to assess the usefulness of regression, three criteria are proposed:

1. The method corrects forecast errors.
2. The method can cope with several highly-correlated predictors which may give rise to multicollinearity.
3. The corrected forecast features the variability of the observation at all lead times, or, in a weaker form, the corrected forecast has the correct variability at long lead times. This condition is necessary in the context of ensemble forecasts in order to get a sufficient spread at long lead times.

2.1 Criterion (i): forecast errors

The corrected forecast should be better than the uncorrected forecast V_1 in the sense that the mean square error between the observation and the corrected forecast (using infinitely large sample sizes) is lower than or equal to the one of the uncorrected forecast.

2.2 Criterion (ii): multicollinearity

Multicollinearity is often encountered when trying to regress a certain variable using highly correlated predictors. In that case, the regression relation may perform well on the training data, but applied to independent data it will give rise to wild and unrealistic results. In that sense, multicollinearity is a form of overfitting, which here is not a consequence of the abundance of regression parameters. Heuristically, multicollinearity can be understood as follows: if two predictors V_1 and V_2 are the same up to a small noise term, ordinary linear regression may yield a predictand which is very close to the training data. However, since V_1 and V_2 are nearly identical, several linear combinations of V_1 and V_2 may exist that are close to the measurement data, the closest of which may involve large and, therefore, unrealistic regression coefficients. Generally, the variances of such estimated coefficients are large and the regression is, thus, very sensitive and unstable with respect to independent data.

Several approaches exist which overcome the multicollinearity problem. One method is to perform a selection

Table 1. Assessment of linear regression methods (rows) based on the criteria (columns) introduced in Sect. 2. A plus sign in brackets means that there exist methods to fulfil this criterion but they are not presented here.

| | crit. (i) | crit. (ii) | weak crit. (iii) | strong crit. (iii) |
|---------|-----------|------------|------------------|--------------------|
| OLS | + | – | – | – |
| TDTR | + | + | + | – |
| TLS | – | – (+) | – | – |
| EVMOS | + | + | + | + |
| GM | – | – (+) | + | + |
| EREG I | + | – (+) | – | – |
| EREG II | – | – (+) | + | + |

and leave out some variables. Another method consists of eliminating the lowest singular value of the matrix which is to be inverted (as used in principal component regression). A third method is called Tikhonov regularization or ridge regression and, from a Bayesian point of view, uses prior information to constrain the regression coefficients.

2.3 Criterion (iii): climatological features at long lead time

It is well known that applying ordinary least-square regression to forecast data amounts to a corrected forecast which converges to the climatological mean for long lead times (Wilks, 2006). This stems from the fact that, due to the chaotic nature of the atmosphere, the correlations between forecast and measurement data irrevocably vanish at long lead time (Vannitsem and Nicolis, 2008). This convergence feature is unrealistic and we can instead try to produce a corrected forecast which has a meaningful climatological variability. Such climatological variability is for most systems well known and usually measured as the variance of the available measurement data, σ_X^2 . Therefore, criterion (iii) states that, in addition to the correct mean, the variability of a good corrected forecast σ_{X_C} should equal the climatological variability of the measurement data:

$$\sigma_{X_C}(t) = \sigma_X(t), \tag{4}$$

or at least convergence towards the climatological variability (weaker constraint):

$$\sigma_{X_C}(t) \rightarrow \sigma_X(t), \tag{5}$$

for long lead times.

Several approaches were already proposed to overcome the lack of variability, including the introduction of an artificial noise term (Unger et al., 2009; Glahn et al., 2009). Let us now introduce the different regression methods compared in the present work. We assess their validity using the aforementioned criteria in Table 1.

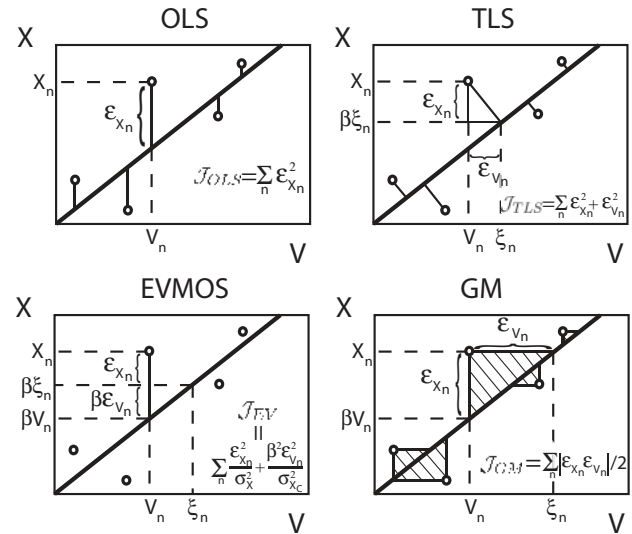


Fig. 1. Illustration of the cost functions \mathcal{J} associated with four different regression methods (OLS, TLS, EVMOS and GM) in the case of one predictor. The dots are data points which associate each forecast point V_n to a corresponding measured value X_n . The black line is the regression line $X_C = \beta V$.

3 Regression methods

3.1 Ordinary least-square (OLS)

Ordinary least-square (OLS) is the most well-known method of linear fitting. It is implicit in OLS regression that there are no forecast uncertainties, but only measurement errors or, in other words, one assumes that $\xi_p = V_p$. The OLS cost function \mathcal{J}_{OLS} is the mean square error of the corrected forecast (see also Fig. 1):

$$\mathcal{J}_{OLS}(\beta) = \langle \epsilon_X^2 \rangle = \langle (X - X_C)^2 \rangle. \tag{6}$$

Here $\langle \cdot \rangle$ is the statistical average over all forecasts. Minimization of Eq. (6) with respect to β yields the well-known solution (Casella and Berger, 1990):

$$\beta_{OLS} = (\mathbf{V}^T \mathbf{V})^{-1} \mathbf{V}^T \mathbf{X}. \tag{7}$$

From Eq. (7) it readily follows that the variance of the corrected forecast variable is:

$$\sigma_{X_C}^2 = \langle X_C^2 \rangle = \langle X X_C \rangle. \tag{8}$$

Therefore, after long lead times, when the correlations $\langle X V_p \rangle$ between the observation and the forecast variable vanishes, the variance of the predictand X_C also vanishes and, therefore, criterion (iii) is not fulfilled. For the MSE this implies:

$$\text{MSE(OLS)} = \langle (X - X_C)^2 \rangle = \sigma_X^2 - \langle X X_C \rangle. \tag{9}$$

At long lead times, the MSE is, therefore, σ_X^2 .

By construction OLS fulfils the requirement of criterion (i). However, apart from criteria (iii), OLS also fails to satisfy (ii) since it cannot cope with multicollinearity, hence, according to our criteria, OLS is not the best method for ensemble regression.

3.2 Time-Dependent Tikhonov Regularization (TDTR)

A well-known problem of the OLS regression is that the variance of the regression parameters are very large in case of multicollinearity (Golub and Van Loan, 1996). To overcome this problem, it is the custom to bias the estimates of the regression coefficients using Tikhonov regularization, also called ridge regression. We present here a new time-dependent Tikhonov regularization (TDTR) method for post-processing ensemble forecasts. The generalized TR approach works as follows: instead of minimizing \mathcal{J}_{OLS} , a constraint is added for the values of the regression coefficients $\boldsymbol{\beta}$ in order to fall within a certain range of a constant value $\boldsymbol{\beta}^0$. More specifically, we demand that $\sum_p (\beta_p - \beta_p^0)^2$ is small. The way to implement such a restriction is by introducing the positive Lagrange multiplier $\gamma(t)$ (with t being time) and minimizing the cost function:

$$\mathcal{J}_{TDTR}(\boldsymbol{\beta}) = \langle (X - X_C)^2 \rangle + \gamma(t) \sum_p (\beta_p - \beta_p^0)^2. \quad (10)$$

The solution is (Björck, 1996):

$$\boldsymbol{\beta}_{TDTR} = \left(\mathbf{V}^T \mathbf{V} + \gamma(t) \mathbf{I} \right)^{-1} \left(\mathbf{V}^T \mathbf{X} + \gamma(t) \boldsymbol{\beta}^0 \right). \quad (11)$$

Here \mathbf{I} is the unit matrix. Using this solution, the variability of the corrected forecast is found to be:

$$\sigma_{X_C}^2 = \langle X X_C \rangle + \gamma(t) \sum_p \beta_p (\beta_p^0 - \beta_p). \quad (12)$$

For the mean square error, on the other hand, one gets:

$$\text{MSE}(\text{TDTR}) = \sigma_X^2 - \langle X X_C \rangle - \gamma(t) \sum_p \beta_p (\beta_p^0 - \beta_p), \quad (13)$$

Using these results, a Tikhonov regularization method can be developed in such a way as to fulfil criteria (ii) and the weak criterion (iii) as specified in Sect. 2. From the inspection of the cost function \mathcal{J}_{TDTR} , it is clear that for large γ the regression coefficient $\boldsymbol{\beta}$ is forced to converge to $\boldsymbol{\beta}^0$. The latter can be chosen in such a way that the corrected forecast has the variability of the measurement data and, thus, satisfies the weak version of criterion (iii). The way to do that is first by Taylor-expanding Eq. (11) up to the first order in $1/\gamma(t)$ (with large $\gamma(t)$) which gives a value of $\boldsymbol{\beta} \approx \boldsymbol{\beta}^0 - \mathbf{V}^T \mathbf{V} \boldsymbol{\beta}^0 / \gamma(t)$, which in turn leads to

$$\sigma_{X_C}(t \rightarrow \infty) = \sigma_{X_C}^0 \quad \text{and} \quad \text{MSE}(t \rightarrow \infty) = \sigma_X^2 + \sigma_{X_C}^0, \quad (14)$$

where $X_C^0 = \sum_p \beta_p^0 V_p$. Note that Eqs. (14) are independent of $\gamma(t)$. Now, if the variability of the predictors is known at long lead time (one predictor would suffice in fact), $\boldsymbol{\beta}^0$ can be chosen in such a way as to satisfy weak criterion (iii): $\sigma_{X_C}(t \rightarrow \infty) = \sigma_X$.

If we also want a scheme able to cope with multicollinearity, or equivalently to fulfil criterion (ii), care must be taken that γ is positive and nonzero but still small at short lead times.

The choice of the time-dependent function $\gamma(t)$ is arbitrary but the cross-over time when γ goes over from being small to being large should preferentially be chosen as a function of the correlations between forecast variables and measurement data. The function used here is:

$$\gamma(t) = \gamma_0 \exp \left\{ \frac{1}{\Delta} \left(\frac{|\text{AC}(0)|}{|\text{AC}(t)|} - 1 \right) \right\}. \quad (15)$$

Here γ_0 is a small positive scalar and AC is an anomaly correlation $\text{AC}(t) = \sum_p \langle X V_p \rangle / (\sigma_X \sigma_{V_p})$ (Van den Dool, 2006). The constant Δ is a tolerance percentage of correlation loss in the sense that, if the anomaly correlation $\text{AC}(t)$ decreases by an amount Δ from its value at time zero, the corrected forecast will become strongly biased towards the solution X_C^0 . Note that, at time zero, $\gamma(0) = \gamma_0$. We choose in our simulations $\gamma = 10^{-4}$ and $\Delta = 0.5\%$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_{EV}$ (as will be defined later in Eq. (24) for the EVMOS technique).

One may come up with choices for γ different from Eq. (15). For example, we can choose $\gamma(t) = \gamma_0 e^{t/\tau}$ where γ_0 is again a small positive constant and τ is a constant which characterises the time when the correlations between the forecast and the measurement start to vanish strongly. In the same line, another candidate could be the threshold function $\gamma(t) = \gamma_0 + \Theta(t - \tau) / \gamma_0$ with Θ the Heaviside function.

Finally, the new TDTR method fulfils criteria (i), (ii) and weak criterion (iii). Note that it is possible to tune the Tikhonov regularization scheme in such way (by means of γ and $\boldsymbol{\beta}^0$) as to fulfil also criterion (iii). However, in that case, one must bias the regression coefficients towards other coefficients on which no information is available, so these must be chosen with some arbitrariness.

3.3 Total least-square (TLS)

The total least-square (TLS) method was introduced as a method to correct OLS to take into account the errors in the forecast model (Golub and Van Loan, 1996). Therefore, the cost function to be minimized is a function of both ε_X and ε_V of Eqs. (2) and (3) (Van Huffel and Vandewalle, 1991):

$$\mathcal{J}_{TLS}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \left\langle \sum_{p=1}^P w_p^2 (V_p - \xi_p)^2 + \left(X - \sum_{p=1}^P \xi_p \beta_p \right)^2 \right\rangle. \quad (16)$$

Here the weight factors w_p are constants which do not depend on the coefficients ξ_p or β_p . After minimizing Eq. (16)

with respect to the variables ξ_p , one gets a cost function which only depends on the values β_p :

$$\mathcal{J}_{\text{TLS}}(\boldsymbol{\beta}) = \frac{\langle (X - X_C)^2 \rangle}{1 + \sum_p (\beta_p/w_p)^2}. \quad (17)$$

After minimization with respect to β_p , one gets an exact solution for the regression problem (Van Huffel and Vandewalle, 1991):

$$\boldsymbol{\beta} = (\mathbf{V}^T \mathbf{V} - \mu^2 \mathbf{W}^{-2})^{-1} \mathbf{V}^T \mathbf{X}, \quad (18)$$

where μ is the lowest singular value of the composite matrix $[\mathbf{V}; \mathbf{X}]$ and $\mathbf{W} = \text{diag}(w_1, \dots, w_p)$. The variability of the corrected forecast is then given by:

$$\sigma_{X_C}^2 = \langle X X_C \rangle + (\sigma_X^2 - \langle X X_C \rangle) \left(\sum_{p_1} (\beta_{p_1}/w_{p_1})^2 \right), \quad (19)$$

and:

$$\text{MSE}(\text{TLS}) = \left(1 + \sum_{p_1} (\beta_{p_1}/w_{p_1})^2 \right) (\sigma_X^2 - \langle X X_C \rangle). \quad (20)$$

Two weaknesses of TLS may be pointed out. First, from Eq. (16), it is clear that, since the different predictors may have different physical units, appropriate estimates of the weight factors w_p prior to regression are indispensable. A second weakness of TLS lies in the fact that the regression estimates become meaningless once a predictor V_p is uncorrelated with the observation X . For instance, in the case of regression with one predictor one finds that $\beta \propto 1/\langle V X \rangle$ which diverges as the correlation vanishes. It follows also that TLS can sometimes fail in satisfying criterion (i). Note that methods exist for TLS to deal with multicollinearity (Van Huffel and Vandewalle, 1991). However, we will not address them here. For the numerical analysis in Sect. 4 we take $w_p = 1$ for all predictor indices p . TLS then minimizes the sum of distances between the data points and the regression line (see Fig. 1).

3.4 Error-in-Variable method (EVMOS)

Recently a new regression method, called EVMOS was introduced for post-processing ensemble forecasts (Vannitsem, 2009). The cost function \mathcal{J}_{EV} takes into account the sum of errors in the forecast variables $\sum_p \beta_p \varepsilon_{V,p}$ and the errors in the measurement variable ε_X (see Fig. 1 for a visual interpretation):

$$\mathcal{J}_{\text{EV}}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \left\langle \left(\frac{X_C - \sum_p \xi_p \beta_p}{\sigma_{X_C}} \right)^2 + \left(\frac{X - \sum_p \xi_p \beta_p}{\sigma_X} \right)^2 \right\rangle. \quad (21)$$

Minimization of Eq. (21) with respect to ξ_p amounts to a cost function which is a function of $\boldsymbol{\beta}$ only:

$$\mathcal{J}_{\text{EV}}(\boldsymbol{\beta}) = \frac{\langle (X - X_C)^2 \rangle}{\sigma_X^2 + \sigma_{X_C}^2}. \quad (22)$$

Further minimization yields for each predictor index p :

$$\sum_{p_1, p_2} \beta_{p_1} \beta_{p_2} c_{X p_1} c_{X p_2} (\rho_{p_1 p_2} - 2\rho_{p p_2}) = -\sigma_X^2. \quad (23)$$

Here $\rho_{p_1 p_2} = \langle V_{p_1} V_{p_2} \rangle / (c_{X p_1} c_{X p_2})$ and $c_{X p} = \langle V_p X \rangle$. This nonlinear problem has a solution:

$$\beta_{p, \text{EV}} = \frac{\sigma_X}{c_{X p}} \left(\frac{\sum_{p_1} (\rho^{-1})_{p p_1}}{\sqrt{\sum_{p_1, p_2} (\rho^{-1})_{p_2 p_1}}} \right). \quad (24)$$

The solution for β_p for up to two predictors was already given in Vannitsem (2009). Note that we provide the details of the derivation in Appendix A.

Using now the solution Eq. (24), it is straightforward to derive that the variance of the corrected forecast is exactly the same as the one of the measured data:

$$\sigma_{X_C}^2 = \sigma_X^2. \quad (25)$$

This equation is valid at all times and, therefore, satisfies criterion (iii) (see Eq. (4)). The MSE becomes:

$$\text{MSE}(\text{EV}) = 2 \left(\sigma_X^2 - \sigma_X \sqrt{\sum_{p_1, p_2} (\rho^{-1})_{p_2 p_1}} \right). \quad (26)$$

One can also make the EVMOS approach robust against multicollinearity by replacing $V_p \rightarrow V_p + \varepsilon_p$ in Eq. (21) with ε_p a noise term with zero mean and standard deviation γ , the latter being a small positive constant. After averaging out the noise terms, one arrives at a new cost function which is minimized by the solution:

$$\beta_p = \frac{\sigma_X}{c_{X p}} \left(\frac{\sum_{p_1} [(\boldsymbol{\rho} + \gamma \mathbf{I})^{-1}]_{p p_1}}{\sqrt{\sum_{p_1, p_2} [(\boldsymbol{\rho} + \gamma \mathbf{I})^{-1}]_{p_2 p_1}}} \right). \quad (27)$$

In practice γ is very small, but still sufficiently large for the variance of the regression coefficients to be sufficiently small. In conclusion, EVMOS is a method which satisfies all proposed criteria.

3.5 Geometric Mean (GM)

The geometric-mean (GM) method with one predictor was introduced by Teisser (1948) and minimizes a cost function which is a sum of triangular areas $|\varepsilon_X \varepsilon_V|/2$ where ε_X and ε_V are the distances from the observation point to the regression line parallel to the X -axis and the V -axis, respectively (see Fig. 1). Draper and Yang (1997) generalized this method to

incorporate multiple predictors. Here we introduce a different approach for such generalization which is still reducible to a weight-free least square problem and satisfies criterion (iii).

The GM approach is introduced in order to overcome one of the main disadvantages of the methods discussed so far which account for both forecast and observation error: for each observation, the associated penalty in the cost function consists of a *sum* of squared errors ε_X and ε_V , normalized using appropriate weight factors. For TLS these are the constants w_p and for EVMOS, σ_X^{-1} and $\sigma_{X_C}^{-1}$. Therefore, those approaches strongly depend on the weight factors. The GM method with one predictor assists in minimizing the sum of triangular areas where the triangles are formed by connecting the fitted line with the measurement point. We extend this approach to more than one predictor by taking the geometric mean of the triangular areas $|\varepsilon_X \varepsilon_{V,p}|/2$ for each predictor index p . The GM cost function then becomes:

$$\mathcal{J}_{\text{GM}}(\boldsymbol{\beta}) = \left\langle \prod_{p=1}^P \left| (V_p - \xi_p) \left(X - \sum_{p_1=1}^P \xi_{p_1} \beta_{p_1} \right) \right|^{1/P} \right\rangle. \quad (28)$$

In order to obtain the value of ξ_p , a projection into the plane formed by the X -axis and the V_p axis should be performed such that $\xi_p = (X - \sum_{p_1 \neq p} V_{p_1} \beta_{p_1}) / \beta_p$. Substitution leads us to the least-square expression:

$$\mathcal{J}_{\text{GM}}(\boldsymbol{\beta}) = \frac{\langle (X - X_C)^2 \rangle}{\left| \prod_p \beta_p \right|^{1/P}}. \quad (29)$$

Minimization of Eq. (29) with respect to β_p gives:

$$-2P\beta_p c_{Xp} \left(1 - \sum_{p_1} \rho_{pp_1} c_{Xp_1} \beta_{p_1} \right) = \langle (X - X_C)^2 \rangle. \quad (30)$$

The minimization problem can also be solved by an iterative numerical method such as the one explained in the Appendix B.

From Eq. (30), one can derive that the variance of the predictand is the same as the one of the measurement variable:

$$\sigma_{X_C}^2 = \sigma_X^2. \quad (31)$$

As was true for the EVMOS approach, this equation is valid at all times and GM, thus, satisfies criterion (iii). Also, the MSE after optimization satisfies:

$$\text{MSE} = 2(\sigma_X^2 - \langle X X_C \rangle). \quad (32)$$

In case of one predictor, the solution to Eq. (30) yields $\beta = \sigma_X / \sigma_V$ which makes it fully equivalent to the EVMOS approach. Thus, EVMOS with one predictor can be considered the minimization solution to a triangle problem. Note also that, in case of one predictor, the GM problem reduces to the one of Draper and Yang (1997).

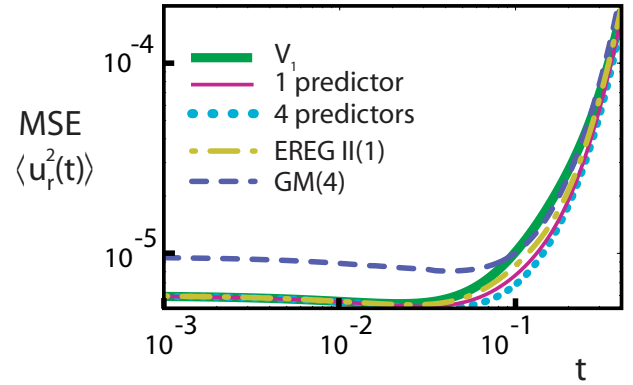


Fig. 2. The short-time mean square error (MSE) evolution for the Lorenz model with a small model error ($\delta r = 2.5 \times 10^{-3}$) and with 5×10^5 ensembles of 500 members. The lines indicated with “1 predictor” and “4 predictors” apply to all regression methods except for GM and EREG II. The numbers indicated after GM and EREG II are the predictor numbers.

3.6 Best-member regression (EREG I and EREG II)

Recently a new approach of ensemble regression was proposed by Unger et al. (2009). The authors show that, if all ensemble members are equally apt at being the best, that is, the closest to reality, and, if a linear relationship exists between the best member and the real data, then the regression coefficient of the OLS can be found using the ensemble mean instead of each ensemble member separately.

Consider a measured path of variable X for which modelling has resulted in an ensemble of K uncorrected forecasts $V_1^k(t)$ ($k = 1, \dots, K$). The ensemble consists of model runs with the same or different models, starting with slightly different initial conditions. We define F now as the average over the ensemble members of the uncorrected forecast, or, $F = \sum_k V_1^k / K$. In order to calculate the regression coefficient, we apply OLS using F instead of V_1 as the model predictor. Minimization of the OLS cost function yields:

$$\beta_{\text{EREG}} = \frac{\langle F X \rangle}{\langle F^2 \rangle}, \quad (33)$$

These regression coefficients are then applied to each ensemble member to yield the best member or EREG I predictand:

$$X_{C,I} = \beta_{\text{EREG}} V_1. \quad (34)$$

From Eq. (33) it is clear that, as is valid for OLS, the variance of the corrected forecast vanishes for long lead times and EREG I will, therefore, not satisfy criterion (iii). However, as was proven in Unger et al. (2009), the damping towards zero of this variance is slower than in the case of OLS. In addition, the authors define an EREG II forecast $X_{C,II}$ which accounts for the lack of spread around the regressed value $X_{C,I}$ by adding artificially a Gaussian noise term with

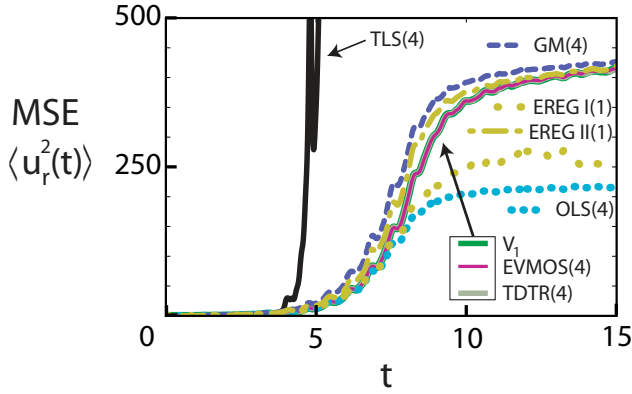


Fig. 3. The long lead time mean square error (MSE) evolution for Lorenz model with *small* model error ($\delta r = 2.5 \times 10^{-3}$) and with 5×10^5 ensembles of 500 members. The number after the regression method denotes the predictor number.

variance $\sigma_X^2 - \sigma_{X_{C,I}}^2$, satisfying criterion (iii). However, such random noise may destroy physically-relevant statistical information of the error statistics. As will be shown in the next section, at intermediate time-scales EREG II may have a MSE which is larger than the one of the uncorrected forecast. Note that the use of more than one predictor can be straightforwardly implemented in the EREG methods as for OLS.

4 Numerical results

We address the usefulness of the different regression methods in the context of a low-order system by focussing on the statistical features of the associated error distributions. We use the well-known Lorenz 1963 model describing thermal convection:

$$\dot{x} = \sigma(-x + y), \tag{35a}$$

$$\dot{y} = rx - y - xz, \tag{35b}$$

$$\dot{z} = xy - bz. \tag{35c}$$

Here the dot denotes the derivative with respect to time, x is the rate of convective turnover, and y and z quantify the horizontal and vertical temperature variation, respectively. The parameter set is fixed to $(\sigma, r, b) = (10, 28, 8/3)$ such that the reference system exhibits chaotic behaviour. The model differs from reality by introducing a model error which we take to be the positive bias δr to the (reduced) Rayleigh number r . For the *small* model-error experiment a bias $\delta r = 2.5 \times 10^{-3}$ is introduced while for the *large* model error experiment a bias $\delta r = 10^{-2}$ is used.

The numerical scheme is integrated using a second-order Runge-Kutta method. An ensemble is constructed by adding at time zero an unbiased Gaussian noise with standard deviation 10^{-3} to all variables x , y and z . In the experiments,

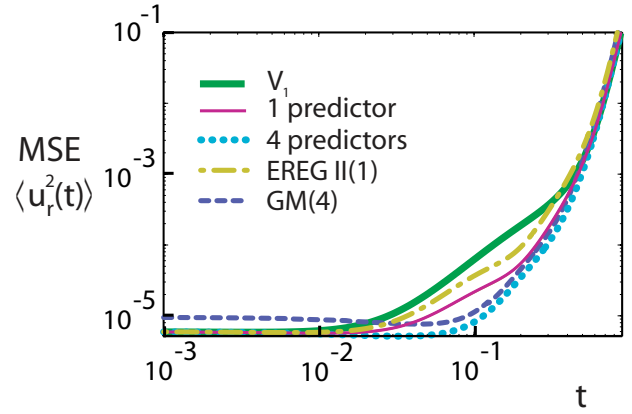


Fig. 4. The short-time mean square error (MSE) evolution for the Lorenz model with *large* model error ($\delta r = 10^{-2}$) and with 5×10^5 ensembles of 500 members. The lines indicated with “1 predictor” and “4 predictors” apply to all regression methods except for GM and EREG II.

we typically use ensembles of 500 members and averaging is typically done over 50 000 points on the attractor. The training and verification of the regression method is performed using two independent datasets, both of the same size.

Originally ensemble forecasts were mainly designed for medium and long range lead times. Nowadays there is a growing interest in using this technique at shorter time scales to provide uncertainty information for short-range forecasting (few hours up to one or two days, Iversen et al., 2010). In the following, we, therefore, present results for the different timescales in order to provide a global picture of the different corrections the post-processing could provide.

We study the errors by probing the statistical properties of one of the following error variables:

$$u_x = x - x_C, u_y = y - y_C, u_z = z - z_C, \text{ and}$$

$$u_r = \sqrt{u_x^2 + u_y^2 + u_z^2}.$$

Here the index C refers to the corrected variable. The predictors for x_C are the variables $(V_1, V_2, V_3, V_4) = (x, y, z, yz)$, generated by different forecasts with model and initial-condition errors and the ones for y_C are obtained by performing the shift $x \rightarrow y \rightarrow z \rightarrow x$; applying again this shift gives the predictors for z_C . In the plots the number of predictors are indicated in brackets next to the regression method.

4.1 Small model error

Let us consider first a small model error $\delta r = 2.5 \times 10^{-3}$ (in the sense that it is comparable to the initial-condition error). Figure 2 shows the time evolution of the MSE of different regression variables at short lead times. Except for a small time window, the MSE of GM(4) is larger than the one of the uncorrected forecast V_1 . This is also the case for the MSE of the EREG II method, but only visible at the intermediate

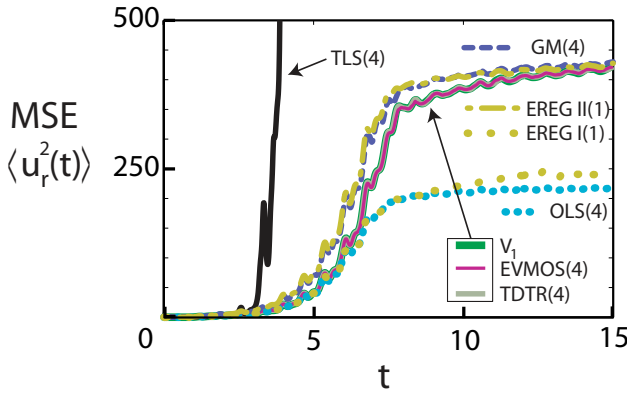


Fig. 5. The mean square error (MSE) evolution for the Lorenz model with *large* model error ($\delta r = 10^{-2}$) for long lead times and with 5×10^5 ensembles of 500 members.

lead times. For all other methods (OLS, TDTR, TLS, EVMOS and EREG I) the results depend only on the number of predictors used and are, therefore, bundled by the lines “1 predictor” and “4 predictors”. Their resulting MSE is always smaller than the one of the uncorrected forecast and provides a substantial forecast improvement only at the smallest timescales where the number and the choice of predictors are important.

GM with one predictor is exactly the same as EVMOS with one predictor. Hence, if instead of GM(4) we would have used GM(1), the results would be indicated by the “1 predictor”-line in Fig. 2 and, thus, have a lower MSE than GM(4). This suggests that GM is progressively degrading when the number of predictors increases. This behaviour can be explained as follows: assume that we apply GM regression with $P - 1$ predictors and continue by adding a new predictor V_P which is totally uncorrelated to all other predictors and to the observation. Then one expects that the regression coefficient β_P associated with this new predictor equals zero. However, from Eq. (30) it can be shown that $\beta_P \propto (2P \langle V_P^2 \rangle)^{-1/2}$. The fact that β_P is generally nonzero could also be suspected from the cost function (29) which diverges if a regression coefficient vanishes. Therefore, adding an uncorrelated predictor introduces an instability. Note that higher-order moments of the error distribution are not well corrected at short lead times, whatever the regression technique used.

At the intermediate lead times there is a fast increase of the MSE as a direct consequence of the chaotic nature of the system (Lorenz, 1963). The timescale involved is determined by the inverse of the dominant Lyapunov exponent. At these lead times no improvements with respect to the original forecast variable V_1 , are achieved by any regression method and the same is true for higher moments of the error distribution. Whereas OLS, TDTR, TLS, EVMOS and EREG I give a MSE equal to the one of the uncorrected forecast, GM(4)

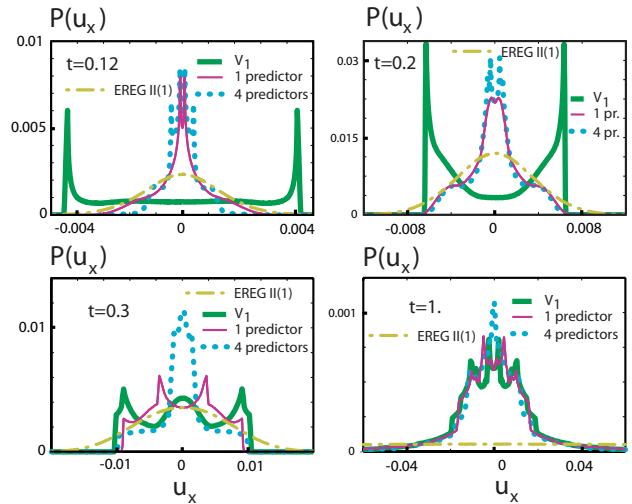


Fig. 6. The distributions of the error variable u_x of the Lorenz model with large model error ($\delta r = 10^{-2}$) and *without* initial condition error, at $t = 0.12, 0.2, 0.3, 1$. The lines indicated with “1 predictor” and “4 predictors” apply to all regression methods except for GM and EREG II. The results are generated using 2×10^6 ensembles.

and the EREG II yield a MSE which is larger. Therefore, GM(4) and EREG II are not well-suited for use at intermediate lead times.

Figure 3 shows the evolution of the MSE for long lead times when the errors become large. Large differences between the regression methods are visible. As mentioned before, the MSE of TLS gives unrealistic results once the correlations between the observation X and the predictors vanish. According to weak criterion (iii), the MSE of the EVMOS, TDTR, EREG II and GM(4) forecasts converges to the correct value $2\sigma_X^2$ at long lead times, but before the asymptotic saturation, the MSE of GM(4) and EREG II is still larger than the one of the uncorrected forecast. The MSE of OLS and EREG I, on the other hand, is too low by a factor of two as the variance of the corrected forecast vanishes.

At the intermediate times, a fast increase is present for all moments of the error distribution, giving rise to a power-law error distribution $P(u_x)$ for large values of the error such that $P(u_x) \propto u_x^{-\nu}$ with some positive scalar ν . Note that a similar behaviour is also present for $P(u_r)$. These power tails are not affected by the different regression methods.

4.2 Large Model Error

The MSE for $\delta r = 10^{-2}$ is plotted in Fig. 4 as a function of time for short and intermediate lead times. It is seen that all regression methods, except GM(4), provide substantial corrections to the model error at short lead times.

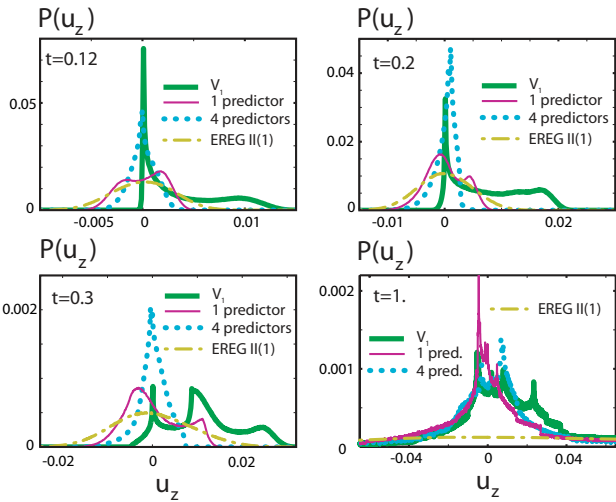


Fig. 7. The distributions of the error variable u_z of the Lorenz model with large model error ($\delta r = 10^{-2}$) and *without* initial condition error, at $t = 0.12, 0.2, 0.3, 1$. The lines indicated with “1 predictor” and “4 predictors” apply to all regression methods except for GM and EREG II. The results are generated using 2×10^6 ensembles.

As for small model errors the amplitude of the corrections obtained by the post-processing progressively decreases. Moreover, both EREG II and GM(4) yield an even higher MSE than the one of the uncorrected forecast. Note that the increase of model errors expands the interval during which GM(4) effectively corrects the forecast. In Fig. 5, we display the MSE for long lead times. As compared to the case of small model errors, the error saturation now sets in earlier and the result for the EREG I forecast is now closer to the one of OLS.

4.3 Evolution of error distribution

Having looked so far at its second moment, we consider now the evolution of the full error distributions of the original and corrected forecast. In Figs. 6 and 7, the error distribution evolution of u_x and u_z in the absence of initial condition errors are plotted. As mentioned earlier, for all methods except for EREG II and GM(4) the quality of regression depends almost solely on the number of predictors.

At time $t = 0.12$ the regression distribution for u_x (Fig. 6) with one predictor is peaked close to the centre. The double-peak feature of the corrected forecast seems to disappear at $t = 0.2$, but appears back again for longer lead time. With four predictors multiple peaks are still present, but the distribution is well centred around zero. At a later time ($t = 1$), all regressed distributions except for EREG II are close to each other, all featuring a multiple-peak structure. The EREG II distribution is by construction a smoother version of the EREG I distribution (here indicated by “1 predictor”) and

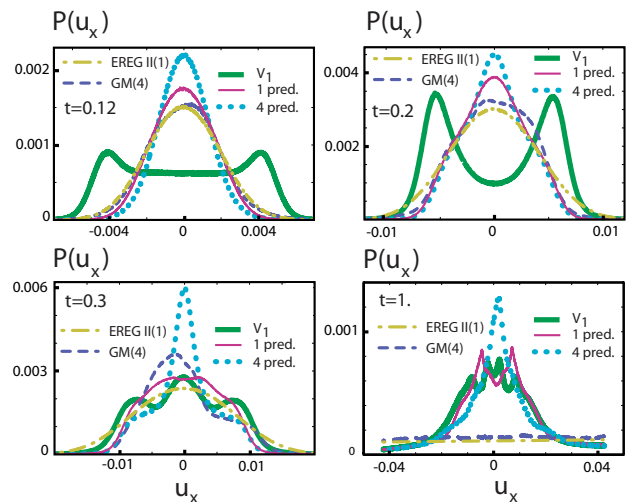


Fig. 8. The distributions of the error variable u_x of the Lorenz model with small model error ($\delta r = 2.5 \times 10^{-3}$) and *with* initial condition errors, at $t = 0.12, 0.2, 0.3, 1$. The lines indicated with “1 predictor” and “4 predictors” apply to all regression methods except for GM and EREG II. The results are generated using 5×10^5 ensembles of 500 members.

tends to be Gaussian-like at all lead times. Such broadening leads to a loss of the statistical information contained in the error distribution. At short times EREG II features a smaller MSE than the one of the uncorrected forecast. However, at longer lead times ($t > 0.2$) the broadening of the distribution produces errors with magnitudes larger than the one present in the uncorrected forecast. At time $t = 1$ the error distribution of EREG II is almost flat with much larger values than the uncorrected forecast. Note that similar results are obtained for other magnitudes of the model error and for the distributions of u_y .

The probability distributions of u_z in Fig. 7 obtained with one predictor removes, to a great extent, the systematic bias of u_z . The “4 predictor” case, on the other hand, also reduces the variance of the error distribution.

Figure 8 displays the results using the same model configuration but *with* initial-condition errors. It is clear that the sharp peaks present in the uncorrected forecast of Fig. 6 are now strongly smoothed out, but their positions are well preserved. The double-peak structure of the distributions after regression, however, seems to have disappeared. As a result of the chaotic nature of the system, the error distributions at time $t = 1$ with and without initial condition errors are very much alike.

In case of nonzero initial-condition errors the qualitative difference between the EREG I and EREG II error distributions are small at the shortest three lead times. This is due to the fact that, by definition, the error distribution of the EREG II method is a smoothed version of the EREG I distribution, which here is a Gaussian-like distribution. However, at long

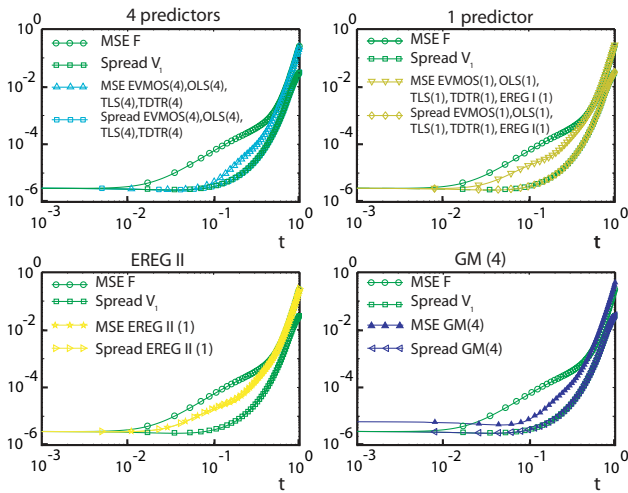


Fig. 9. Ensemble variance (spread) and mean square error of the ensemble mean of the uncorrected forecast V_1 against corrected forecasts produced by different regression methods as a function of time, generated using the Lorenz model with large model error ($\delta r = 10^{-2}$) and averaged over 5×10^5 ensembles of 500 members each. Note that the EREG II ensemble variance and mean square error of the ensemble mean are identical.

lead times, due to the large noise variance used to generate the EREG II forecast, the MSE of EREG II is well beyond the uncorrected forecast, leading to an almost flat unrealistic distribution.

5 Ensemble features

As pointed out in the Introduction, the main reason for looking at alternative linear post-processing is to investigate the possibility of post-processing *ensemble* forecasts. One important reason for the use of ensemble forecasts is that it provides one with an estimate of the forecast uncertainty. In this section we explore, using numerical experiments, how the relationship between the ensemble spread and forecast accuracy is affected by post-processing. We compare first the average ensemble error with the average ensemble spreads, and we proceed by considering the relation between error and spread of each ensemble separately.

One requirement for a good ensemble forecast is to have a mean square error of the ensemble mean equal to the ensemble variance (e.g. Leutbecher and Palmer, 2008). In Fig. 9 the error of the ensemble mean is compared with the ensemble variance. The corresponding quantities for the uncorrected forecast are also displayed using green symbols. Except for EREG II, regression does not affect the ensemble variance of the uncorrected forecast at short lead times. The error dynamics of the ensemble mean, on the other hand, is similar to the mean square error evolution of all ensemble members as shown in Fig. 4. Note that such regressions give rise to en-

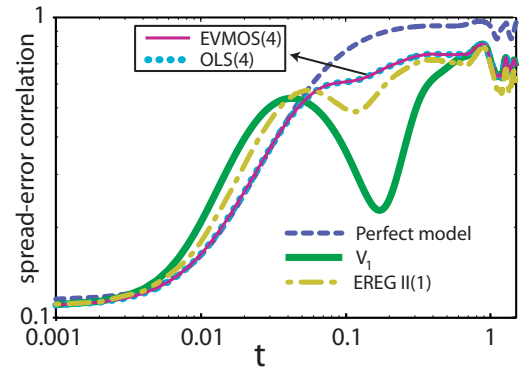


Fig. 10. Log-log plot of the ensemble spread-error correlation against time for short time scales. The spread is the square root of the ensemble variance and the error is the ensemble average of u_r . The results are generated using the Lorenz model with large model error ($\delta r = 10^{-2}$) and averaged over 2×10^5 ensembles of 500 members each. The perfect-model result is obtained without model error.

sembles which remain underdispersive except for the EREG II ensemble. Due to the unbiased noise used to construct the EREG II ensemble, the MSE of the ensemble mean of EREG I and EREG II are identical. The gain of ensemble variance without loss of accuracy of the ensemble mean constitutes the most interesting feature of EREG II but it is obtained at the expense of an increase of the overall ensemble member error as shown in Fig. 4 and a broadening of the error distribution (e.g. Figs. 6 and 8). In agreement with our previous results, at long lead times when the errors are saturated, the average ensemble variance converges to the error of the ensemble mean for EREG II, EVMOS, TDTR and GM ensembles.

We study now in what sense the ensemble spread can be considered a measure of the actual error and how it is affected by post-processing. Figure 10 shows the Pearson correlation between ensemble spread and ensemble error. The ensemble spread refers to the square root of the ensemble variance and the error is the root-mean-square error u_r over all ensemble members. In Grit and Mass (2007) these were found adequate variables for a correlation study where it was also pointed out that the correlations are intimately related to the variance of the ensemble spread. Figure 10 focuses on the shortest timescale for the perfect-model ensemble (using no model error and the correct initial-condition distribution), the model ensemble with large model error (V_1) and the corrected ensemble using OLS, EVMOS and EREG II. Initially the correlation is low (around 0.1) due to a small standard deviation of ensemble spread, which is about ten times smaller than the average ensemble spread. One can even observe that the perfect model error-spread correlation is initially smaller than all others. At a timescale of about 0.1, however, post-processing is found to substantially increase

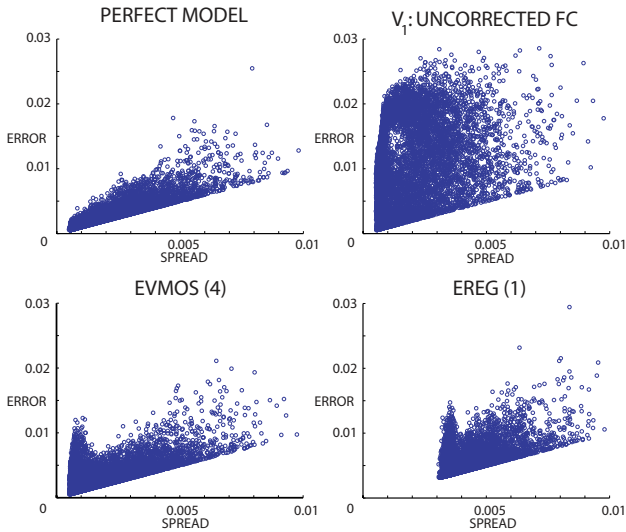


Fig. 11. Scatterplot of spread against error of 10 000 ensembles at $t = 0.2$. The spread is the square root of the ensemble variance and the error is the ensemble average of u_r . The results are generated using the Lorenz model with large model error ($\delta r = 10^{-2}$) using ensembles of 500 members each. The perfect-model result is obtained without model error.

the correlation. We illustrate this in Fig. 11 with scatterplots of spread against error for 10 000 ensembles at $t = 0.2$. The strong spread-error correlation for the perfect model is obvious from a clustering of dots along the diagonal. On the contrary, the imperfect-model ensembles are strongly dispersed. The post-processing procedures are capable of strongly reducing the errors such that a great deal of spread-error correlation is recovered. Note that the addition of random noise of the EREG II method amounts to a shift of the ensemble cloud along the diagonal.

The correlations at intermediate and long timescales are plotted in Fig. 12. At intermediate timescales the spread-error correlation is large as the standard deviation of ensemble spread is now on average ten times larger than the average ensemble spread. Even though the spread-error correlation of the uncorrected and the EVMOS ensembles are approximately equal to the one of EREG II, the associated errors and spreads strongly differ. This is illustrated in Fig. 13 by scatterplots at $t = 2$. The uncorrected and EVMOS ensemble clouds are much alike but the EREG II cloud is shifted along the diagonal, a transformation which preserves the linear correlation. Note also the enlarged scale of the EREG II plot. Finally, due to the error saturation, a progressive correlation decrease sets in for all ensembles at lead times $t = 5$ (see Fig. 12). Remarkably, the OLS and EREG II correlations are distinctly smaller than the ones of the uncorrected and EVMOS ensembles. At $t = 15$, the variance of ensemble spread for all except the EREG II ensembles is still significant as suggested in Fig. 14.

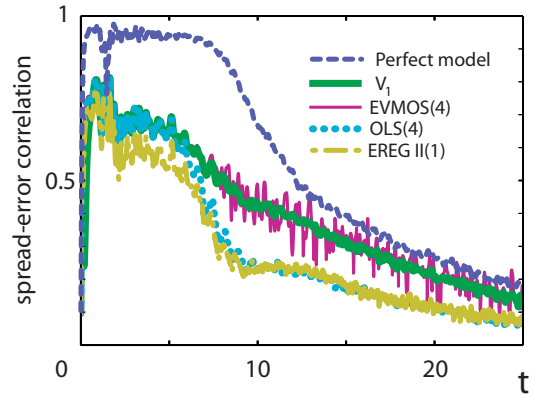


Fig. 12. The spread-error correlation against time for intermediate and long times. The spread is the square root of the ensemble variance and the error is the ensemble average of u_r . The results are generated using the Lorenz model with large model error ($\delta r = 10^{-2}$) and averaged over 2×10^5 ensembles of 500 members each.

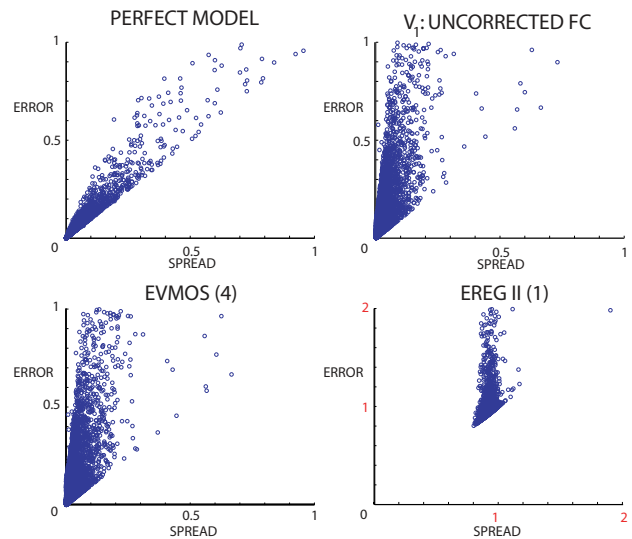


Fig. 13. Scatterplot of spread against error of 10 000 ensembles at $t = 2$. The spread is the square root of the ensemble variance and the error is the ensemble average of u_r . The results are generated using the Lorenz model with large model error ($\delta r = 10^{-2}$) using ensembles of 500 members each. The perfect-model result is obtained without model error. Note the scale of the EREG II plot which has a scale which is twice as large.

Figure 15 displays the smoothing effect of EREG II by showing the time evolution of the x -variable of ten ensemble members (full lines), along with the measured trajectory (dashed line). All members are initialized around the initial position $(x, y, z) = (-11.84, -4.484, 38)$ on the attractor, very close to two bifurcating trajectories and using an initial condition spread of 10^{-1} . The ensuing separation of the

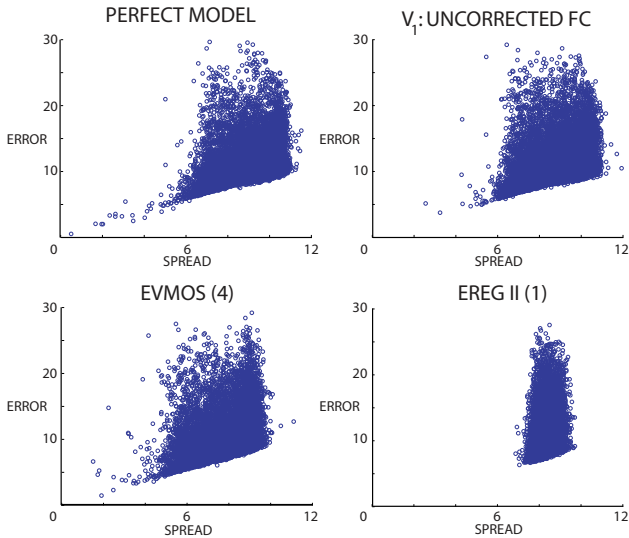


Fig. 14. Scatterplot of spread against error of 10 000 ensembles at $t = 15$. The spread is the square root of the ensemble variance and the error is the ensemble average of u_r . The results are generated using the Lorenz model with large model error ($\delta r = 10^{-2}$) using ensembles of 500 members each. The perfect-model result is obtained without model error.

ensemble members into two separate regions is well reproduced by EVMOS (as well as TDTR), but due to the random noise this feature is no longer present for lead times $t > 5$ in the EREG II forecast. The absence of the bimodality of the EREG II distribution for the variable x at lead time $t = 5.29$ is clearly seen in Fig. 16.

6 Conclusions

Several linear regression methods have been tested in the context of post-processing of (ensemble) forecasts: classical linear regression, total least-square regression, Tikhonov regularization, error-in-variable regression, geometric mean regression and best-member regression. These approaches were evaluated based on three criteria (see Table 1): a correction of the forecast error, the ability to cope with multicollinearity and the reproduction of the observed variability.

The regression schemes have been tested in the context of the low-order Lorenz 1963 system by introducing both model and initial-condition errors. Three timescales may be distinguished. First, for short lead times, strong error improvements and an increase of ensemble spread-error correlations may be obtained in case of large model errors. Except for GM, skill at these timescales does not so much depend on which regression method is applied, but rather on how many and which predictors are selected. Second, at intermediate times, when the error (and all the moments of its distribution) undergoes fast growth, all regression methods, except

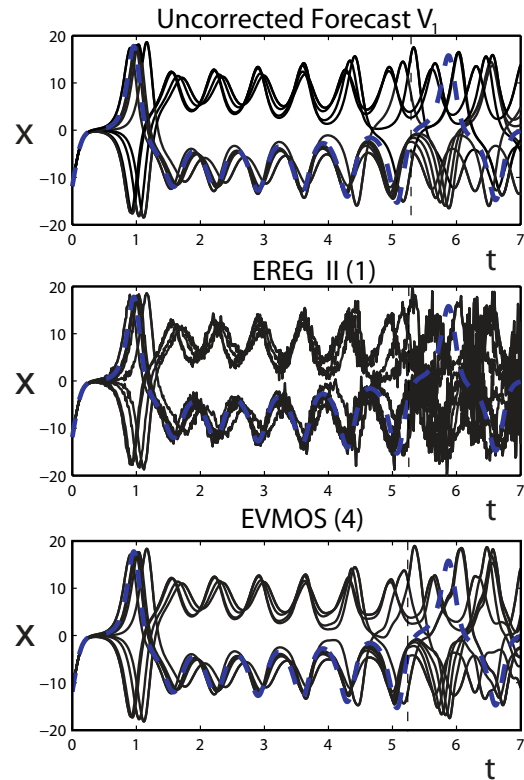


Fig. 15. Time evolution of the variable x for ten ensemble members (full lines) using the Lorenz model generated by the uncorrected forecast V_1 , EREG II and EVMOS(4). The blue dashed line indicated the evolution of the observation. All ensemble members are initialized using an initial condition spread 10^{-1} around the starting point $(x, y, z) = (-11.84, -4.484, 38)$.

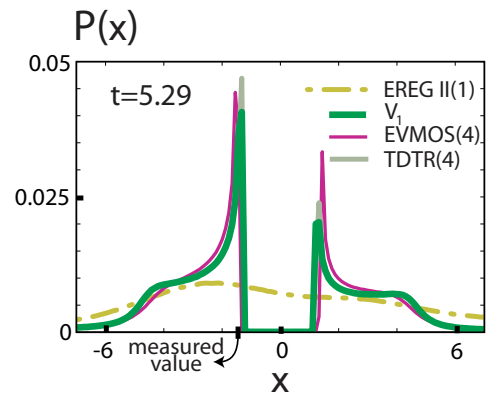


Fig. 16. Probability distributions of the variable x at time $t = 5.29$ for the ensembles generated as specified in Fig. 15. We used 10^7 ensemble members. It is clear that the EREG II distribution is not bimodal as are the others.

for EREG II and GM, yield the same result. EREG II and GM provide less favorable results. Third, for long times, when the correlation between the measurement variable and the predictors is almost zero, strong differences between the regression methods are visible. TLS yields a wild and unphysical forecast; the OLS and EREG I corrected forecasts converge towards climatology.

The EREG II method has the benefit to account for the lack of variability which is featured in the EREG I method and, thus, satisfies criterion (iii). Also it provides a mean ensemble spread which is very close to the MSE of the ensemble mean, a property often required in an operational context. However, all this is done by construction: random noise is added *a posteriori* to the EREG I forecast. This implies that some essential statistical information of the underlying physics is lost, such as the specific multiple-peak structure of the error distribution and the non-Gaussian nature of the probability distribution of the variables themselves. A similar behaviour is encountered in an operational context when EVMOS was compared with the non-homogeneous Gaussian regression method, the latter smoothing out the multimodal structure of the forecast (Vannitsem and Hagedorn, 2011). Also, at long lead times EREG II, as well as OLS and EREG I, has a reduced spread-error correlation as compared to the uncorrected and the EVMOS forecast.

Another technique exists in the literature based on the potential relation between the different observables effectively measured (Perfect Prog, Klein et al., 1959; Wilks, 2006). This approach does not suffer from the convergence towards climatology like OLS. However, the correction obtained with this technique is useless for sufficiently small forecast error. We have performed a preliminary exploration of this aspect by applying Perfect Prog for the Lorenz system. For each variable X, Y, Z, we have built a Perfect Prog relationship based on the two other variables of our reference system. In this case, Perfect Prog becomes useful when errors reach values of the order of 1/5 of the saturation error variance shown in Fig. 5.

The techniques as presented here can be extended to combine multimodel forecasts (e.g. Peña and Van den Dool, 2008). A straightforward way would be to use the different forecasts as predictors. Regression corresponds to a method for “weighting” the different models. However, since many such models may contain the same information, one must be sure that the regression method is able to cope with multicollinearity. TDTR and EVMOS can fulfil this requirement.

Appendix A

A1 Derivation of EVMOS solution

In this Appendix, we provide additional calculations concerning the EVMOS method as introduced in Sect. 3.4. The EVMOS cost function \mathcal{J}_{EV} of Eq. (21) can be rewritten as a

sum over all forecasts:

$$\mathcal{J}_{EV}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_n \left[\left(\frac{\sum_p (V_{np} - \xi_{np}) \beta_p}{\sigma_{X_C}} \right)^2 + \left(\frac{X_n - \sum_p \xi_{np} \beta_p}{\sigma_X} \right)^2 \right]. \quad (\text{A1})$$

First, we minimize with respect to ξ_{np} . This yields:

$$\sum_p \beta_p \xi_{np} = \frac{\sigma_X^2 \sum_p \beta_p V_{np} + \sigma_{X_C}^2 X_n}{\sigma_X^2 + \sigma_{X_C}^2}, \quad (\text{A2})$$

and substitution in Eq. (A1) gives:

$$\mathcal{J}_{EV}(\boldsymbol{\beta}) = \frac{\sum_n \left(X_n - \sum_p \beta_p V_{np} \right)^2}{\sigma_X^2 + \sigma_{X_C}^2}.$$

The variance of the predictand $\sigma_{X_C}^2$ can be written as $\sum_{p_1, p_2} \beta_{p_1} \langle V_{p_1} V_{p_2} \rangle \beta_{p_2}$. Minimization with respect to β_t then gives:

$$\begin{aligned} & \left(\sum_p \langle V_t V_p \rangle \beta_p - \langle V_t X \rangle \right) \left(\sigma_X^2 + \sum_{p_1, p_2} \beta_{p_1} \langle V_{p_1} V_{p_2} \rangle \beta_{p_2} \right) \\ &= \left(\sum_p \langle V_t V_p \rangle \beta_p \right) \left(\sigma_X^2 + \sum_{p_1, p_2} \beta_{p_1} \langle V_{p_1} V_{p_2} \rangle \beta_{p_2} \right) \\ & \quad - 2 \sum_p \langle X V_p \rangle \beta_p. \end{aligned} \quad (\text{A3})$$

We introduce now the vector $\bar{\boldsymbol{\beta}}$ with components $\bar{\beta}_p = \beta_p c_{Xp}$ where $c_{Xp} = \langle X V_p \rangle$, and the matrix $\boldsymbol{\rho}$ with components $\rho_{p_1 p_2} = \langle V_{p_1} V_{p_2} \rangle / (c_{Xp_1} c_{Xp_2})$. After some calculation, Eq. (A3) reduces to:

$$\bar{\boldsymbol{\beta}}^T \boldsymbol{\rho} \bar{\boldsymbol{\beta}} - 2 \left(\mathbf{1}^T \cdot \bar{\boldsymbol{\beta}} \right) \bar{\boldsymbol{\beta}}^T \boldsymbol{\rho} = -\sigma_X^2 \mathbf{1}, \quad (\text{A4})$$

with $\mathbf{1}$ a vector with all its P components equal to one. One can check that the following solution satisfies Eq. (A4):

$$\bar{\boldsymbol{\beta}} = \frac{\sigma_X (\boldsymbol{\rho}^{-1} \cdot \mathbf{1})}{\sqrt{\mathbf{1}^T \cdot \boldsymbol{\rho}^{-1} \cdot \mathbf{1}}}, \quad (\text{A5})$$

which is identical to the solution given in Eq. (24). Note that the expression under the square-root sign is always positive since $\boldsymbol{\rho}$ is a correlation matrix and is, therefore, positive definite, as well as its inverse.

Appendix B

B1 Numerical method for nonlinear GM regression

The GM cost function Eq. (29) can be minimized using the iterative Gauss-Newton methods for least-square problems (see for example Björck, 1996). First of all, the cost function can be written as a function of the matrix \mathbf{r} :

$$\mathcal{J}_{GM}(\boldsymbol{\beta}) = \mathbf{r}^T \cdot \mathbf{r}, \quad (\text{B1})$$

where $\mathbf{r}(\boldsymbol{\beta}) = (\mathbf{X} - \mathbf{V}\boldsymbol{\beta})/\mathcal{N}$ and $\mathcal{N}(\boldsymbol{\beta}) = \left| \prod_p \beta_p \right|^{1/2P}$. Then given at step k the values \mathbf{r}^k and the regression coefficients $\boldsymbol{\beta}^k$, the coefficients at step $k+1$ by assuming small changes are searched $\Delta\boldsymbol{\beta} = \boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^k$. Therefore, $r_n(\boldsymbol{\beta})$ at step $k+1$ becomes:

$$\mathbf{r}(\boldsymbol{\beta}^{k+1}) \approx \mathbf{r}(\boldsymbol{\beta}^k) + \mathbf{J}^k \Delta\boldsymbol{\beta}, \quad (\text{B2})$$

where the matrix $\mathbf{J}_{np} = \partial r_n / \partial \beta_p$, evaluated at step k . Substituting Eq. (B2) into the cost function Eq. (B1) at time step $k+1$ and expanding to second order yields:

$$\mathcal{J}_{\text{GM}}^{k+1}(\boldsymbol{\beta}) \approx \mathbf{r}^T \cdot \mathbf{r} + \Delta\boldsymbol{\beta}^T \mathbf{J}^T \mathbf{J} \Delta\boldsymbol{\beta} + 2\mathbf{r}^T \mathbf{J} \Delta\boldsymbol{\beta}, \quad (\text{B3})$$

where again, all is evaluated at time step k . Minimization with respect to $\Delta\boldsymbol{\beta}$ gives:

$$\Delta\boldsymbol{\beta} = -(\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \mathbf{r}. \quad (\text{B4})$$

So far, the derivation was general. Substituting now the cost function of the GM approach, one gets:

$$\Delta\beta_p = \beta_p (\mathbf{A}^{-1} \mathbf{B})_p, \quad (\text{B5})$$

where:

$$\begin{aligned} \mathbf{A}_{p_1 p_2} &= \widehat{\mathbf{C}}_{p_1 p_2} + \frac{(\bar{\boldsymbol{\beta}} - \widehat{\mathbf{C}} \cdot \mathbf{1})_{p_1} + (\bar{\boldsymbol{\beta}} - \widehat{\mathbf{C}} \cdot \mathbf{1})_{p_2}}{2P} \\ &\quad + \frac{\sigma_X^2 + \mathbf{1}^T \cdot \widehat{\mathbf{C}} \cdot \mathbf{1} - 2(\mathbf{1}^T \cdot \bar{\boldsymbol{\beta}})}{(2P)^2}, \\ \mathbf{B} &= \frac{\bar{\boldsymbol{\beta}} - \widehat{\mathbf{C}} \cdot \mathbf{1}}{2P} + \frac{\sigma_X^2 + \mathbf{1}^T \cdot \widehat{\mathbf{C}} \cdot \mathbf{1} - 2(\mathbf{1}^T \cdot \bar{\boldsymbol{\beta}})}{(2P)^2}, \end{aligned}$$

where we introduced the notations $\widehat{\mathbf{C}}_{p_1 p_2} = \rho_{p_1 p_2} \bar{\beta}_{p_1} \bar{\beta}_{p_2}$, $\bar{\beta}_p = \beta_p c_{Xp}$ and $\mathbf{1}$ is a vector containing P times the scalar one. We may now calculate $\boldsymbol{\beta}^{k+1} = \Delta\boldsymbol{\beta} + \boldsymbol{\beta}^k$ and continue this procedure to time step $k+1$ until convergence is reached.

Acknowledgements. We thank Anastasios Tsonis and an anonymous reviewer for their constructive comments. This work is supported by the Belgian Science Policy Office under contract MO/34/020.

Edited by: J. Duan

Reviewed by: A. Tsonis and another anonymous referee

References

- Björck, A.: Numerical methods for least-square problems, SIAM, Philadelphia, 408 pp., 1996.
- Casella, G. and Berger, R. L.: Statistical Inference, Brooks Cole Publishing, 661 pp., 1989.
- Draper, N. R. and Yang, Y.: Generalization of the geometric-mean functional relationship, *Comput. Stat. Data An.*, 23, 355–372, doi:10.1016/S0167-9473(96)00037-0, 1997.
- Glahn, B., Peroutka, M., Wiedenfeld, J., Wagner, J., Zylstra, G., Schuknecht, B., and Jackson, B.: MOS Uncertainty Estimates in an Ensemble Framework, *Mon. Weather Rev.*, 137, 246–268, 2009.
- Golub, G. H. and Van Loan, C. F.: Matrix computations, 794 pp., 1996.
- Grimit, E. P. and Mass, C. F.: Measuring the ensemble spread-error relationship with a probabilistic approach: stochastic ensemble results, *Mon. Weather Rev.*, 135, 203–221, 2007.
- Iversen, T., Deckmyn, A., Santos, C., Sattler, K., Bremnes, J. B., Feddersen, H., and Frogner, I.-L.: Evaluation of “GLAMEPS” – a proposed multi-model EPS for short range forecasting, *Tellus A*, in press, doi:10.1111/j.1600-0870.2010.00507.x, 2010.
- Klein, W. H., Lewis, B. M., and Enger, I.: Objective prediction of five-day mean temperatures during winter, *J. Meteor.*, 16, 672–682, 1959.
- Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, *J. Comput. Phys.*, 227, 3515–3539, 2008.
- Lorenz, E. N.: Deterministic non-periodic flows, *J. Atmos. Sci.*, 20, 130–141, 1963.
- Peña, M. and Van den Dool, H.: Consolidation of Multimodel Forecasts by Ridge Regression: Application to Pacific Sea Surface Temperature, *J. Climate*, 21, 6521–6538, 2008.
- Teisser, G.: La relation d’allométrie sa signification statistique et biologique, *Biometrics*, 4, 14, 1948.
- Unger, D. A., van den Dool, H., O’Lenic, E., and Collins, D.: Ensemble Regression, *Mon. Weather Rev.*, 137, 2365–2379, 2009.
- Van Huffel, S. and Vandewalle, J.: The total least-square problem: Computational aspects and analysis, SIAM, Philadelphia, 300 pp., 1991.
- Van den Dool, H. M.: Empirical Methods in Short-Term Climate Prediction, Oxford University Press, 215 pp., 2006.
- Vannitsem, S.: Dynamical Properties of MOS Forecasts: Analysis of the ECMWF Operational Forecasting System, *Weather Forecast.*, 23, 1032–1043, 2008.
- Vannitsem, S.: A unified linear Model Output Statistics scheme for both deterministic and ensemble forecasts, *Q. J. Roy. Meteorol. Soc.*, 135, 1801–1815, 2009.
- Vannitsem, S. and Hagedorn, R.: Ensemble forecast post-processing over Belgium: Comparison of deterministic-like and ensemble regression methods, *Meteorol. Appl.*, 18, 1, 94–104, March 2011.
- Vannitsem, S. and Nicolis, C.: Dynamical Properties of Model Output Statistics Forecasts, *Mon. Weather Rev.*, 136, 405–419, 2008.
- Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic Press, New York, 468 pp., 2006.